Knowledge Driven Graph Evolution (KDGE)

Federico Tomasi[1], Margherita Squillario[1], Annalisa Barla[1]

[1] Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS),
University of Genoa, Genoa, I-16146, Italy
federico.tomasi@dibris.unige.it, margherita.squillario@unige.it,
annalisa.barla@unige.it

Abstract

Understanding complex biological phenomena is a difficult task, considering the interplay among molecular variables. Common approaches rely on data to build a robust statistical model. Usually, a set of variables which are able to characterise the biological conditions under analysis is identified, and then enrichment analysis is used *a posteriori* in order to give a functional assessment of selected variables. In this case, prior knowledge is used to validate the result instead of leading the analysis from the start. In this work, we make explicit use of prior knowledge, using signalling pathways during the learning phase. After, a network (or graph) of inferred pairwise interactions is extracted using mutual information scores in the different biological conditions, to assess interaction evolution of biological variables.

1 Motivation

Understanding molecular variables which discriminate a set of clinical outcomes is not sufficient to completely and exhaustively explain the molecular mechanisms that lead to different biological conditions. In fact, the interplay and interaction of molecular variables plays a key role in characterising a clinical outcome. Usually, prior knowledge on the interplay of such variables (*i.e.*, involvement in a common pathway) is used after data analysis by means of functional characterisation of those variables. This approach has some drawbacks. Misguided variable selection and classification procedures may include many false hits, or, even worse, may exclude variables potentially relevant in the biological context under analysis. In particular, this may happen when variables are not important *per se*, but they may be deemed biologically relevant if considered within a molecular module.

Prior biological knowledge can be effectively used when learning the statistical model underlying the data, guaranteeing the non-exclusion of variables that are biologically relevant to the analysed disease. Selecting the relevant modules only partially shed light on the phenomenon under study. In particular, the interplay among variables within the same module may change when considering different biological conditions. In this context, network (or graph) reconstruction methods are powerful tools that allow to depict the evolution of the module.

Here, we describe a novel machine learning pipeline, which identifies meaningful signalling pathways, reconstructs the corresponding networks and quantitatively assesses pathways evolution under different conditions. We present the results on a RNA-seq breast invasive carcinoma dataset.

2 Materials and Methods

2.1 Data

We use a RNA-seq dataset of breast invasive carcinoma (BRCA), consisting of 20,501 gene expression measures of 822 patients. The dataset belongs to the TCGA Pancan project, publicly available online[1]. Samples belong to eight different classes, based on their clinical information. Patients are identified based on estrogen receptor positive (ER+) or negative (ER-). Each group is further divided into four classes, based on the lymph node involvement of their disease (N0–N3).

Information on 1859 (human) signalling pathways is extracted from Reactome [2], a curated and peer-reviewed pathway database.

2.2 Pathway selection

We consider a multiclass problem managed using a one-vs-rest scheme, that is, a binary classification problem is fit for each label. For each pathway, a learning machine consisting of a regularised logistic regression model with ℓ_1 or ℓ_2 penalty (based on the dimensionality of the pathway) is iteratively fit, validated and tested via a model assessment framework based on Monte Carlo cross-validation [1]. The output of this procedure is an estimation of repeated learning and test scores. The robustness of the system is also tested against chance, by means of randomly permuting the labels and re-evaluate the procedure. Finally, the resampled distribution is compared to the random distribution, and a p-value is estimated. The choice of a sparsity-enforcing classifier (via the ℓ_1 penalty) allows to have a list of most selected variables which are important to discriminate output classes. For pathways which contained less than 100 variables, no feature selection step was employed, and the ℓ_2 penalty was used. The final outcome is predicted as the consensus among all different estimators.

2.3 Network inference and assessment

For each class of the patients, a network is extracted using a pairwise mutual information score, used in many algorithms for network inference. In particular, we use ARACNE [7] on the expression of genes selected by the estimators. Two nodes are said to interact based on their pairwise mutual information. To avoid an over-representation of the network, only edges with a weight higher than the average in the network are retained.

Network differences are quantitatively assessed using the normalised Hamming network distance [3].

3 Results

Among the 1859 pathways we selected those which obtained the best predictive accuracy and the lowest p-value, indicating the significance of the result and the module under analysis.

[1] `https://www.synapse.org/#!Synapse:syn1461151`

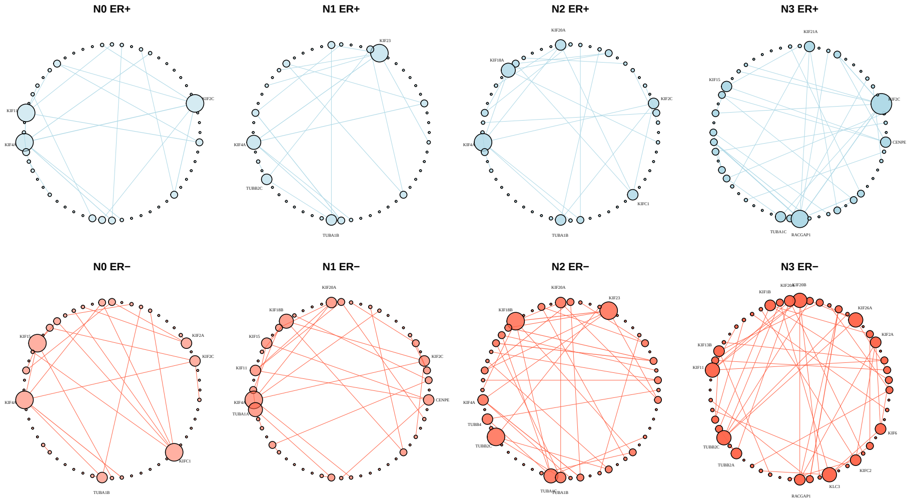

Figure 1: Network inference on 56 selected variables of Kinesins pathway (R-HSA-983189) for BRCA-affected patients based on the estrogen receptor and lymph node involvement.

We report results for the Kinesins pathway (R-HSA-983189), as it harbours groups of genes significantly associated to both lymph node involvement and response to estrogens. Figure 1 shows an increasing number of interactions between kinesins from N0 to N3 lymph node involvement stage, both for ER+ (top row) and ER- (bottom row). On average, graphs associated to ER- have a higher number of interactions with respect to graphs associated to ER+. The distance between networks under different biological conditions were assessed using the Hamming network distance. In Figure 2, a hierarchical clustering algorithm built on such precomputed distances highlights how the network differences are related to the increasing of lymph node involvement considering both ER+ and ER- groups.

Kinesin are microtubule-based motor proteins that mediate diverse functions within the cell, including the transport of vesicles, organelles, chromosomes and

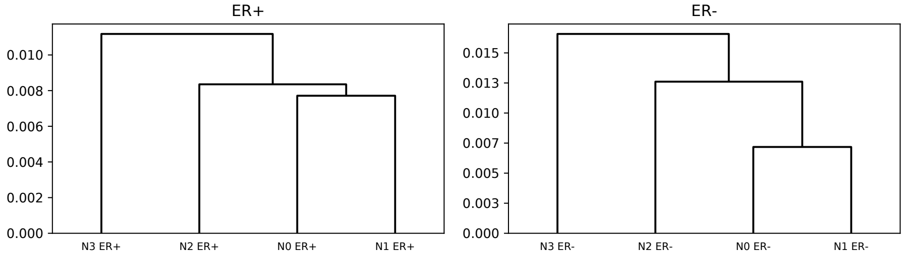

Figure 2: Network distances based on estrogen receptor and lymph node involvement. Both dendrograms show that the aggregation of consecutive stages occurs sequentially.

protein complexes, as well as the movement of microtubules. Recently, these proteins have been given a central role in the regulation of mitotic events and potential targets breast cancer therapy, among others [4]. This pathway contains important genes which have been associated to breast cancer, such as KIF20A, KIF20B, TUBB2A and TUBB2C (TUBB4B). The up-regulation of KIF20A, together with its transcription factor FOXM1, is significantly associated with poor survival and with Paclitaxel action and resistance, a chemotherapy medication used to treat breast cancer and other tumour types [5]. Although not as largely characterised, KIF20B share with KIF20A the interactor FOXM1, as experimentally tested. Results on the Kinesin pathway suggests that KIF20B could be involved in the malignant progression of breast cancer. This hypothesis is supported by the co-occurrence of KIF20A and KIF20B in the N3 ER- graph. TUBB2A and TUBB2C genes share an annotation status similar to KIF20A and KIF20B. This means that TUBB2A is more annotated than TUBB2C and the up-regulation of TUBB2A, as well as KIF20A, is known to be significantly associated to Paclitaxel action and resistance [6]. While poorly annotated, TUBB2C is known to interact with TUBB2A. Results suggests that TUBB2C could be involved in the Paclitaxel action and resistance, hypothesis supported by the co-occurrence of TUBB2A and TUBB2C and in the N3 ER- graph.

4 Conclusion

Knowledge Driven Graph Evolution (KDGE) is a novel method for uncovering structures under development or evolution among nodes of a biological pathway. Also, the use of network distances allows to quantitatively address those differences. The method allows to reconstruct networks of gene interactions based on different biological conditions. Networks extracted for each stage of the disease allow to understand how the evolution of the disease affects the interplay among variables involved in specific pathways.

References

[1] Matteo Barbieri et al. “PALLADIO: a parallel framework for robust variable selection in high-dimensional data”. In: *Proceedings of the 6th Workshop on Python for High-Performance and Scientific Computing*. IEEE Press. 2016, pp. 19–26.

[2] Antonio Fabregat et al. “The reactome pathway knowledgebase”. In: *Nucleic acids research* 44.D1 (2016), pp. D481–D487.

[3] Richard W Hamming. “Error detecting and error correcting codes”. In: *Bell Labs Technical Journal* 29.2 (1950), pp. 147–160.

[4] Phillip Kaestner and Holger Bastians. “Mitotic drug targets”. In: *Journal of cellular biochemistry* 111.2 (2010), pp. 258–265.

[5] P Khongkow et al. “Paclitaxel targets FOXM1 to regulate KIF20A in mitotic catastrophe and breast cancer paclitaxel resistance”. In: *Oncogene* 35.8 (2016), pp. 990–1002.

[6] Luis J Leandro-Garcia et al. “Regulatory polymorphisms in β-tubulin IIa are associated with paclitaxel-induced peripheral neuropathy”. In: *Clinical Cancer Research* 18.16 (2012), pp. 4441–4448.

[7] Adam A Margolin et al. “ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context”. In: *BMC bioinformatics* 7.1 (2006), S7.